4-1-2007

Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies

Miriam K. Konkel
Center for BioModular Multi-Scale Systems

Jianxin Wang
Roswell Park Cancer Institute

Ping Liang
Roswell Park Cancer Institute

Mark A. Batzer
Center for BioModular Multi-Scale Systems

Recommended Citation

Konkel, M., Wang, J., Liang, P., & Batzer, M. (2007). Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. *Gene, 390* (1-2), 28-38. https://doi.org/10.1016/j.gene.2006.07.040

Available online at www.sciencedirect.com
ScienceDirect
Gene 390 (2007) 28–38
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies

Miriam K. Konkel [a,1], Jianxin Wang [b,1], Ping Liang [b], Mark A. Batzer [a,*]

[a] *Department of Biological Sciences, Biological Computation and Visualization Center, Center for BioModular Multi-Scale Systems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA*
[b] *Department of Cancer Genetics, Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, NY 14263, USA*

Received 9 June 2006; received in revised form 18 July 2006; accepted 26 July 2006
Available online 30 August 2006

Abstract

Mobile elements represent a relatively new class of markers for the study of human evolution. Long interspersed elements (LINEs) belong to a group of retrotransposons comprising approximately 21% of the human genome. Young LINE-1 (L1) elements that have integrated recently into the human genome can be polymorphic for insertion presence/absence in different human populations at particular chromosomal locations. To identify putative novel L1 insertion polymorphisms, we computationally compared two draft assemblies of the whole human genome (Public and Celera Human Genome assemblies). We identified a total of 148 potential polymorphic L1 insertion loci, among which 73 were candidates for novel polymorphic loci. Based on additional analyses we selected 34 loci for further experimental studies. PCR-based assays and DNA sequence analysis were performed for these 34 loci in 80 unrelated individuals from four diverse human populations: African-American, Asian, Caucasian, and South American. All but two of the selected loci were confirmed as polymorphic in our human population panel. Approximately 47% of the analyzed loci integrated into other repetitive elements, most commonly older L1s. One of the insertions was accompanied by a BC200 sequence. Collectively, these mobile elements represent a valuable source of genomic polymorphism for the study of human population genetics. Our results also suggest that the exhaustive identification of L1 insertion polymorphisms is far from complete, and new whole genome sequences are valuable sources for finding novel retrotransposon insertion polymorphisms.

Keywords: Retrotransposons; Insertion polymorphisms; Mobile elements

1. Introduction

The decoding and assembly of the complete human genome have contributed significantly to the understanding of human evolution and the structure of the human genome (Lander et al., 2001). The human sequence exhibits several interesting (and partly unexpected) characteristics; *e.g.*, coding sequences comprise less than 5% of the human genome, whereas repetitive elements account at least for 50% of the genome (Lander et al., 2001). The majority of repetitive elements are represented by retrotransposons (~42% of the genome). This group can be divided into two groups: Long Terminal Repeat (LTR) or retrovirus-like; and non-LTR retrotransposons, especially *L*ong *IN*terspersed *E*lement*s* (LINEs). LINE-1 (L1), which emerged about 120 million years ago, represents the youngest and only currently active LINE element in primates (Smit et al., 1995). Roughly 17% or 520,000 L1 copies occupy the human genome (Brouha et al., 2003; Lander et al., 2001; Smit, 1996). However, the majority of these elements are 5-truncated, 5-inverted, internally rearranged, or mutated, and therefore no longer able to retrotranspose (Kazazian and Moran, 1998; Ostertag and Kazazian, 2001).

While a full-length consensus human L1 sequence is roughly 6 kb long (Kazazian and Moran, 1998), the average size of all L1

Abbreviations: LINE-1 or L1, long interspersed element-1; BLAT, BLAST like alignment tool; BLAST, basic local alignment search tool; UTR, untranslated region; ORF, open reading frame; TSD, target site duplication; PHG, Public Human Genome; CHG, Celera Human Genome; LTR, long terminal repeat; Hs, human specific; Ta, transcribed subset a; bp, base pair; EN, endonuclease; RT, reverse transcriptase; PCR, polymerase chain reaction; SNP, single nucleotide polymorphism; chr, chromosome.

* Corresponding author. Tel.: +1 225 578 7102; fax: +1 225 578 7113.
E-mail address: mbatzer@lsu.edu (M.A. Batzer).

[1] These authors contributed equally to this research.

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.07.040

elements in the genome is ~900 bp. The median size for the currently active L1 element in Homo sapiens (L1Hs) is about 1070 bp (Lander et al., 2001). It is estimated that only ~60 to 100 L1 copies in a given genome are capable of retrotransposition; roughly 10% of these active elements are classified as "hot" (highly active, Brouha et al., 2003; Sassaman et al., 1997), with allele specific differences in retrotransposition capability (Lutz et al., 2003; Seleme et al., 2006). The majority of these retrotransposition competent elements belong to the subgroup of L1-Ta (Transcribed subset a) elements (Skowronski et al., 1988), and can be distinguished from pre-Ta elements based on a 3 bp triplet at position 5930–5932 within the 3-UTR-sequence (Dombroski et al., 1991).

Full-length L1s harbor an internal RNA polymerase II promoter in the 5-UTR (UnTranslated Region); have two non-overlapping Open Reading Frames (ORFs), separated by a 63 bp spacer region; and are followed by a 3-UTR region and a poly-A tail (Kazazian, 1998). While ORF1 encodes a 40 kDa RNA-binding protein, ORF2 contains a 150 kDa protein with both endonuclease (EN) and reverse transcriptase (RT) activities (Feng et al., 1996; Jurka, 1997; Mathias et al., 1991). L1s are surrounded by Target Site Duplications (TSDs), which are characterized by a short duplication of locus-specific sequence (Fanning and Singer, 1987).

As first described for *Alu* (short interspersed elements) insertion polymorphisms (Batzer et al., 1996, 1994; Batzer and Deininger, 1991; Deininger and Batzer, 1993, 1999; Perna et al., 1992; Stoneking et al., 1997), a *de novo* insertion of a transposable element creates a new polymorphic locus with unique properties of identity by descent and known ancestral state (Batzer and Deininger, 1991, 2002). Insertion homoplasy – an independent insertion at exactly the same position – can occur, but it is a very rare event with recently integrated *Alu* elements (Hillis, 1999; Ray et al., 2006; Roy-Engel et al., 2002a; Salem et al., 2003b).

Insertion presence/absence polymorphisms have also been described for L1Hs elements. Previous studies of L1-Ta (transcribed subset a) insertions indicated that they have retrotransposed since the origin of our species (Boissinot et al., 2000, 2004; Myers et al., 2002; Ovchinnikov et al., 2001; Sheen et al., 2000). As this group is currently active, it comprises almost all of the *de novo* disease-associated L1 insertions (Kazazian, 1998; Kazazian and Moran, 1998). Pre-Ta elements, likely the ancestors of Ta elements, are also potentially polymorphic, but to a smaller extent (Salem et al., 2003a). There is evidence that some of these elements remain active (Kazazian et al., 1988). Collectively, L1 insertion polymorphisms represent useful markers for population diversity and evolutionary studies (Boissinot et al., 2000, 2004; Myers et al., 2002; Nikaido et al., 1999; Salem et al., 2003a). Also, as shown for *Alu* elements (Ray et al., 2005a), L1 insertion polymorphisms have potential use in forensics.

Thus far, the search for L1 insertion polymorphisms has been limited to either a few individuals screened for insertion polymorphisms, or computational analysis of genomic data (Badge et al., 2003; Bennett et al., 2004; Boissinot et al., 2000, 2004; Buzdin et al., 2003; Myers et al., 2002; Ovchinnikov et al., 2001; Roy-Engel et al., 2001; Salem et al., 2003a, 2000). Thus, the list of polymorphic retrotransposon insertions currently present in the human population is far from complete. With the availability of two almost complete human genome sequences (Human Genome Consortium and Celera), the opportunity arose to compare these two genomes in order to identify LINE insertion polymorphisms. Here, we compare the Public Human Genome (PHG) and Celera Human Genome (CHG) assemblies to identify novel L1 insertion polymorphisms that are differentially inserted in the two genomes.

2. Materials and methods

2.1. Origin of genomic sequences

Two human genomic sequence assemblies were used: the Public Human Genome (PHG) version (Lander et al., 2001), obtained from the UCSC site (May 2004 freeze or hg17) at http://genome.ucsc.edu; and the Celera Human Genome (CHG) from the Celera Discovery System (August 2003 version) through private database subscription (http://cds.celera.com; Venter et al., 2001). Information regarding the sources of DNA used for the creation of the genomic libraries that were sequenced to generate the two genome assemblies can be obtained from the original publications (Lander et al., 2001; Venter et al., 2001). The CHG is based on unconnected scaffolds grouped by chromosome. In order to retrieve the highest accuracy and amount of information, we also analyzed the CHG whole shotgun assembly (WGSA) sequences from GenBank (accessions AADD010000001–AADD01211493, Istrail et al., 2004). All sequences (in Fasta format) were downloaded onto our local bioinformatics server for further analysis.

2.2. In silico identification of L1 insertion polymorphisms

The identification of L1 insertions specific to each genome assembly (*i.e.* candidate L1 insertion polymorphisms) was performed using a method described in Lee et al. (in press). Where Lee et al. compared the PHG with the chimpanzee genome, we have performed this comparison with the CHG. Briefly, the L1 insertions in each assembly were identified by Basic Local Alignment Search Tool (BLAST, Altschul et al., 1990) queries of the genome with the 50 bp sequences preceding the poly-A tail of the L1 consensus sequence, oriented toward the 3-side (Fig. 1). Next, 300 bp sequences at the 3-end of the L1 element, together with 100 bp of flanking sequence immediately downstream of the poly-A tail, were extracted. The ending position of poly-A tails in these L1 sequences was determined using BLAST, with the 50 bp L1 consensus sequence carrying a tract of 100 adenosines as the poly-A tail. The resulting sequences were used as queries for BLAST searches against the CHG. Queries with matches limited to the 100 bp of flanking regions beyond the L1 3-end were collected as candidate L1 insertion polymorphisms from the PHG. These correspond to the pre-integration sites of L1 insertions in the CHG. We then extracted an 800-bp region from the CHG centered at the pre-integration site,

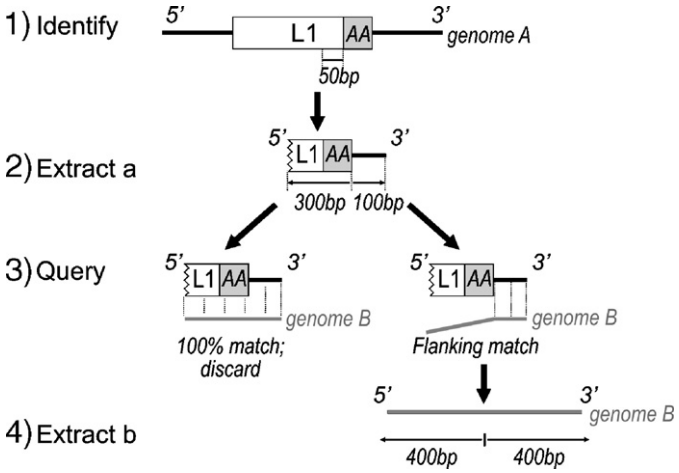

Fig. 1. *In silico* identification of L1 insertion polymorphisms. AA represents poly-A tail. Genome A can be either the PHG or CHG. If genome A is PHG, genome B is CHG; and vice versa. There are four analytic stages. 1) Identify putative polymorphic L1 insertion loci using 50 bp L1 consensus sequence from genome A; 2) extract the 3-end of each L1 element and 100 bp adjacent flanking sequence (genome A); 3) query genome B; discard locus if 100% match between the L1 and flanking sequence (shared insertion); if only flanking sequence matches, continue with step 4; 4) extraction of 400 bp upstream and downstream from point of insertion (genome B). This is followed by extraction of L1 sequence plus 400 bp of flanking on each side (genome A), and alignment of CHG/PHG sequences (not shown).

as well as the L1 insertion together with a 400 bp upstream and downstream flanking sequence from the PHG. To reduce false positives, we retained only those loci with a high degree of identity between the PHG and CHG in the 800 bp flanking regions. The procedure was repeated by reversing the order of the PHG and CHG assembly. All candidate loci were mapped against the Database of Retrotransposon Insertion Polymorphisms (dbRIP, http://falcon.roswellpark.org:9090/; Wang et al., 2006b), to distinguish between novel and previously identified L1 insertion polymorphisms.

To further reduce false positive insertion polymorphisms, for all novel candidate loci, we submitted both the PHG and CHG sequences to BLAT (BLAST Like Alignment Tool, http://genome.ucsc.edu/, Kent, 2002) to determine whether they each mapped to the same genomic location. When loci resided within repetitive sequences, the empty and filled alleles sometimes showed essentially identical matches with their BLAT results, but mapped to different genomic locations. These loci were dropped from the analysis because they represented inappropriate alignments between the PHG and CHG that contained monomorphic or fixed insertions. The remaining sequences were analyzed in RepeatMasker (http://www.repeatmasker.org/, Smit et al., 1996) to identify the class/subclass of the L1 insertions, and to check flanking sequences for repetitive sequences in preparation for designing primers.

2.3. Cell lines and DNA samples

The human cell line HeLa (American Type Culture Collection [ATCC] number CCL2) was used for primer optimizations. HeLa cells were maintained as directed by the source, and DNA extraction was performed with Wizard genomic DNA purification (Promega) per protocol. For our population panel, human DNA samples from European and African American population groups (isolated from peripheral blood lymphocytes) were available from previous studies (Myers et al., 2002; Stoneking et al., 1997). DNA from Asian and South American populations was purchased from the Corriell Institute for Medical Research.

2.4. Primer design

In general, primers were designed using Primer3 (http://cbr-rbc.nrc-cnrc.gc.ca/cgi-bin/primer3_www.cgi, Rozen and Skaletsky, 1998). Each primer sequence was analyzed with BLAT to ensure at least one member of a pair resided in unique sequence. In cases where BLAT reported multiple matches within the genome, those matches were separated sufficiently that the incorrect match would not produce a PCR product. Virtual PCR was also performed for each primer combination to ascertain the size of the product and to further confirm that only one product was likely to be generated. Candidate loci residing within repetitive elements required special consideration. In these cases, BLAT usually showed only a few very closely matching genomic locations. For each of these problematic loci, we aligned the sequences obtained by the BLAT searches and identified point mutations among the sequences which allowed us to design primers (with the 3-end on a point mutation) that would amplify only the desired product. In general, primers were designed within the adjacent 300 bp of flanking sequence. However, some amplicons were slightly larger to facilitate amplification. Longer insertions were genotyped as previously described (Sheen et al., 2000). Sequences of the primers (including the size of PCR products) are available as supplemental data on our website (http://batzerlab.lsu.edu).

2.5. PCR amplification and DNA sequence analysis

In our study, 80 individuals from four different regions (20 Asians, 20 South Americans, 20 African-Americans, and 20 Caucasians) were genotyped using PCR assays. Each locus was analyzed with at least two separate PCR reactions. PCR amplification was performed in 25 μl reactions containing 10 to 25 ng of template DNA; 200 nmol of each oligonucleotide primer; 1.5 mM $MgCl_2$; 50 mM KCl; 10 mM Tris–HCl (pH 8.4); 0.2 mM dNTPs; and 2 U *Taq* DNA polymerase (Myers et al., 2002; Roy-Engel et al., 2002a). Each PCR reaction was performed under the following conditions: initial denaturation step at 94 °C for 150 s, followed by 32 cycles of denaturation at 94 °C for 15 s, 15 s at annealing temperature (specific for each locus), and an extension step at 72 °C for 30 s, followed by a final extension at 72 °C for 3 min. For analysis, 15 μl of each PCR product was fractionated on a 2% agarose gel containing 0.1 μg/ml ethidium bromide in a horizontal gel chamber for 45 min at 180 V, and directly visualized using UV-fluorescence. Prior to full scale analysis, PCR products from each allele of each locus were analyzed by the chain-termination sample sequencing (Sanger et al., 1977) using an Applied Biosystems 3100 automated DNA sequencer to verify their authenticity.

3. Results

3.1. In silico identification of polymorphic LINE-1 insertions and brief characterization

A total of 148 potentially polymorphic L1 insertions were isolated through an *in silico* comparison of the PHG (129 L1s) and the CHG (19 L1s). After comparing these to the dbRIP database, 73 potentially novel polymorphic L1 insertions remained (65 in the PHG, 8 in the CHG). In 38 cases, BLAT analyses identified the loci as false positives. In these cases, the loci were determined to be fixed elements that appeared polymorphic because flanking repetitive elements had led to inappropriate alignments of the two databases. To ensure that our computational rejection of these 38 loci was accurate, we performed a PCR analysis with selected loci. Of the remaining 35 loci, 17 resided in repetitive sequences. Primers were designed for all but one of these 35 loci. One locus was excluded because primer design was not possible. Of the 34 selected insertions, 28 insertions were identified in the PHG, and 6 in the CHG.

3.2. LINE-1 subfamilies of potentially polymorphic loci

All L1 subfamilies were potentially included in the *in silico* analysis to identify polymorphic loci. However, only one older element (L1PA2) qualified for a PCR based analysis; all other candidates were L1Hs subfamily members. Based on diagnostic mutations at positions 5930–5932 in the 3-UTR-sequence of L1s (Dombroski et al., 1991) the majority of the candidate loci (22) were identified as L1-Ta elements (Table 1). Five insertions were L1-pre-Ta elements. Six L1Hs loci could not be further subgrouped, as the inserted L1 sequence was either too short (five loci), or the diagnostic mutations were inconclusive (one locus).

3.3. Confirmation of polymorphic loci

We performed a survey of human genomic diversity associated with all 34 potentially polymorphic insertions by genotyping 80 individuals (160 haploid genomes) from four different geographic regions. Fig. 2 shows two examples from a PCR-based analysis of one population (20 individuals): a truncated L1 element (one PCR required to genotype) and a long L1 insertion (two separate PCRs were necessary for amplification of the filled and empty site).

In this analysis, 32 of 34 loci were confirmed as authentic L1 insertion polymorphisms. Detailed information for these 32 novel verified polymorphic L1 loci has been deposited into dbRIP (Wang et al., 2006b). The two excluded loci were removed for different reasons. In one case, the insertion – an L1Hs-Ta element – was absent in all tested individuals. Several explanations are possible for this finding. This could be caused by an assembly problem, but this appears less likely because the insertion is flanked by unique sequence. Instead, we believe it is likely a result of a very low-frequency element or private insertion polymorphism, as shown previously for some *Alu* elements (Carroll et al., 2001). Another locus, containing an L1PA2 element, consistently showed a heterozygous genotype for all individuals. While it is conceivable this locus is polymorphic, it is very unlikely that all 80 individuals are heterozygous. This PCR pattern is consistent with an insertion that integrated in some type of undefined sequence that is repeated within the genome (*i.e.* segmental duplication). In these cases, the pre-integration site which is repeated throughout the genome generates the empty site product, and the filled site is generated by a single L1 element that is fixed for the L1 element insertion. This is consistent with previous analyses of human *Alu* elements (Batzer et al., 1991).

Table 1
Subclassification of potentially and confirmed polymorphic L1 insertions

L1 subclass	L1Hs-Ta	L1Hs-pre-Ta	L1Hs[a]	L1PA2
Number of loci	22 (70%)	5 (15%)	6 (12%)	1 (3%)
Confirmed loci	21	5	6	0

[a] These elements were too short to be subclassified (insertions < 130 bp), but were identified as L1Hs; and elements which lacked the diagnostic mutation for being clearly characterized as L1Hs-Ta or pre-Ta.

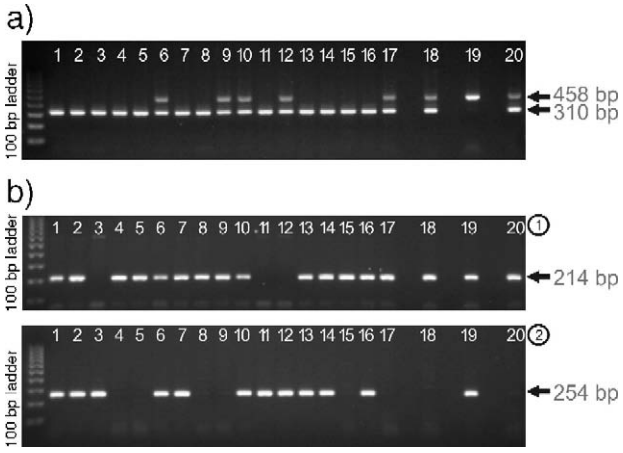

Fig. 2. Examples of two L1 insertion polymorphisms. Agarose gel chromatographs of PCR products derived from an analysis of L1 insertion polymorphisms with short (a) and long (b) lengths are shown. Eighty individuals from four populations (20 individuals each) were assayed for the presence and/or absence of the L1 insertion. a) Polymorphic truncated L1Hs insertion (chr7-1438) in 20 Caucasians. Amplification of the non-insertion (empty) site generates a 310 bp PCR product; amplification of the inserted site (filled) created a PCR product of 458 bp. The chromatograph shows that in tested Caucasians, all three possible genotypes – homozygote insertion present (*e.g.*, I-19), homozygote insertion absent (*e.g.* I-1), heterozygote (*e.g.* I-6) – are present. The majority of the surveyed individuals were homozygous for absence of the L1 insertion. b) Survey of human diversity of a long L1Hs insertion polymorphism (chr1-5654). Here, PCR-results of 20 South Americans are shown. Due to the size of the L1 insertion (4069 bp), two separate PCR amplifications were necessary to genotype the samples. (1) shows the genotyping of the empty alleles with primers in the flanking sequence, creating a PCR product with the size of 214 bp. In a second reaction, an internal, Ta-subfamily specific primer, and the 3-flanking unique sequence primer were used to genotype filled sites (2), the amplification of the filled site generated a 254 bp PCR product. The chromatograph shows that in the tested Caucasians, all three possible genotypes – homozygote insertion present (*e.g.*, I-3), homozygote insertion absent (*e.g.* I-1), heterozygote (*e.g.* I-4) – are present.

3.4. L1 insertion-associated human diversity

For the 32 confirmed novel polymorphic loci, the median allele frequency over all insertion polymorphisms and

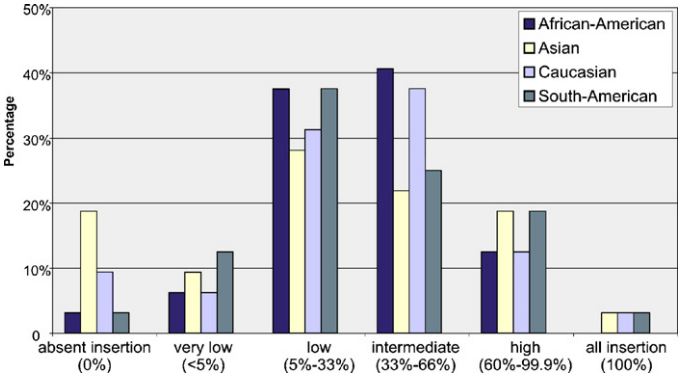

Fig. 3. L1 insertion polymorphism allele frequencies. Here, all confirmed novel polymorphic L1 loci are combined. Insertions are separated by frequency class, ranging from completely absent to always present. In contrast to all other populations, more insertions were absent from the Asian population. Also, Asians were less frequently represented in the group of low to intermediate allele frequencies.

populations was 36% and did not differ considerably between the four populations (35.15% Asian to 37.66% African-American). But the median varied substantially over the four populations, ranging from 22.5% in Asians to 37.5% in African-Americans. To investigate the latter finding, we examined the allele frequency distribution, separated by populations for all loci (Fig. 3). Asians were more often represented both in the group of very low to absent elements, and also in the group with high allele frequency. Asians had the highest frequency of loci with no insertion present (18.75%), followed by Caucasians (9.4%) and African-Americans/South Americans (one locus or 3.1% each).

In order to determine if the newly identified autosomal L1 insertion polymorphisms were in Hardy–Weinberg Equilibrium (HWE) we compared expected genotype frequencies with observed genotype frequency using chi-square tests for goodness of fit. A total of 124 chi-square tests for goodness of fit are possible. However, 14 of the comparisons involved populations that were monomorphic for the presence or absence of the L1 insertion leaving 110 possible tests. A total of 19 deviations from Hardy–Weinberg expectations were observed in the comparisons. Twelve of the deviations were the result of low expected genotype frequencies. Of the remaining seven tests that deviated from HWE, none clustered by population or locus. This deviation is not surprising since a total of 5.5 deviations from HWE would be expected by chance alone at the 5% significance level.

Table 2
Allele frequencies of all polymorphic loci, including the two non-polymorphic insertions

		African-American				Asian				Caucasian				South American			
		Allele freq	−/−	+/−	+/+	Allele freq	−/−	+/−	+/+	Allele freq	−/−	+/−	+/+	Allele freq	−/−	+/−	+/+
Celera	chr1-2182	40.0%	7	10	3	27.5%	10	9	1	30.0%	10	8	2	17.5%	14	5	1
	chr5-8948	95.0%	0	2	18	100.0%	0	0	20	100.0%	0	0	20	100.0%	0	0	20
	chr7-1100	60.0%	4	8	8	85.0%	0	6	14	40.0%	5	14	1	40.0%	8	8	4
	chr8-8390	12.5%	16	3	1	0.0%	20	0	0	7.5%	17	3	0	15.0%	15	4	1
	chr8-8738	25.0%	11	8	1	87.5%	0	5	15	75.0%	1	8	11	70.0%	2	8	10
	chr17-2542	40.0%	4	16	0	22.5%	11	9	0	32.5%	7	13	0	30.0%	8	12	0
Public Human Genome (PHG)	chr1-8645	0.0%	20	0	0	0.0%	20	0	0	35.0%	7	12	1	30.0%	11	6	3
	chr1-5654	70.0%	2	8	10	45.0%	4	14	2	40.0%	9	6	5	37.5%	8	9	3
	chr2-1761	30.0%	9	10	1	15.0%	14	6	0	62.5%	3	9	8	50.0%	5	10	5
	chr2-1958	22.5%	11	9	0	5.0%	18	2	0	5.0%	18	2	0	5.0%	18	2	0
	chr2_8804	30.0%	11	6	3	10.0%	16	4	0	25.0%	12	6	2	27.5%	10	9	1
	chr3_2206	37.5%	8	9	3	20.0%	13	6	1	37.5%	7	11	2	40.0%	7	10	3
	chr3_8301	27.5%	10	9	1	0.0%	20	0	0	22.5%	11	9	0	22.5%	12	7	1
	chr4_1031	45.0%	5	12	3	65.0%	3	8	9	62.5%	3	9	8	75.0%	2	6	12
	chr4_1253	80.0%	0	8	12	77.5%	1	7	12	75.0%	2	6	12	85.0%	2	2	16
	chr4_1398	20.0%	12	8	0	15.0%	14	6	0	0.0%	20	0	0	2.5%	19	1	0
	chr4_1526	45.0%	6	10	4	65.0%	4	6	10	42.5%	6	11	3	47.5%	7	7	6
	chr4_1904	60.0%	3	10	7	35.0%	9	8	3	22.5%	12	7	1	27.5%	12	5	3
	chr4_8124	55.0%	3	12	5	45.0%	2	18	0	50.0%	4	12	4	37.5%	5	15	0
	chr5_1038	12.5%	16	3	1	32.5%	9	9	2	37.5%	6	13	1	22.5%	12	7	1
	chr6_2006	50.0%	4	12	4	67.5%	2	9	9	77.5%	1	7	12	72.5%	0	11	9
	chr6_5758	50.0%	0	20	0	50.0%	0	20	0	50.0%	0	20	0	50.0%	0	20	0
	chr6_8676	10.0%	16	4	0	22.5%	12	7	1	0.0%	20	0	0	2.5%	19	1	0
	chr7_1438	2.5%	19	1	0	0.0%	20	0	0	22.5%	12	7	1	12.5%	15	5	0
	chr7_3216	37.5%	9	7	4	10.0	16	4	0	5.0%	18	2	0	5.0%	18	2	0
	chr7_9047	15.0%	14	6	0	42.5%	6	11	3	35.0%	9	8	3	37.5%	7	11	2
	chr8_7174	67.5%	2	9	9	85.0%	0	6	14	60.0%	3	10	7	67.5%	3	7	10
	chr10_1277	62.5%	3	9	8	52.5%	4	11	5	35.0%	9	8	3	40.0%	7	10	3
	chr11_1218	2.5%	19	1	0	0.0%	20	0	0	10.0%	17	2	1	12.5%	16	3	1
	chr11_2430	60.0%	4	8	8	82.5%	2	3	15	85.0%	1	4	15	85.0%	1	4	15
	chr15_5300	27.5%	11	7	2	0.0%	20	0	0	0.0%	20	0	0	0.0%	20	0	0
	chr16_2695	17.5%	13	7	0	5.0%	18	2	0	7.5%	17	3	0	10.0%	16	4	0
	chrX_1462	45.0%	10	2	8	5.0%	19	0	1	7.5%	18	1	1	7.5%	17	3	0
	chrX_1618	0.0%	20	0	0	0.0%	20	0	0	0.0%	20	0	0	0.0%	20	0	0

−/− insertion absent; +/− heterozygote insertion present/absent; +/+ insertion present; and Allele freq: allele frequency.

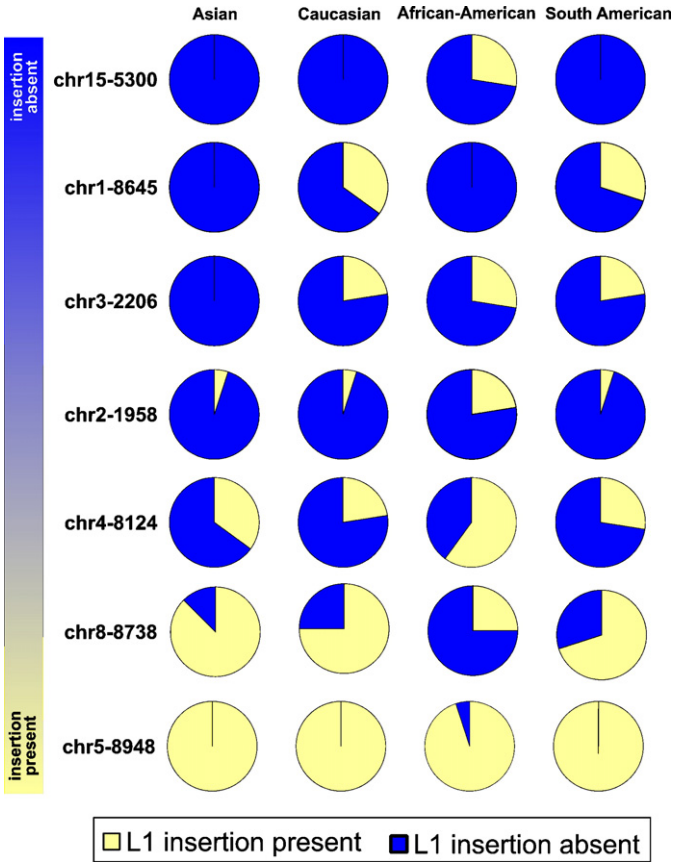

Fig. 4. Selection of allele frequencies over four populations. A selection of insertion polymorphisms is shown. Included in this diversity analysis were 20 individuals from each of the four populations, totaling 80 individuals. Different populations are listed from right to left; different loci (with insertion absent to insertion present), from top to bottom.

Table 2 shows the allele frequencies of each individual polymorphic locus separated by population. Some of the analyzed polymorphisms showed frequency specificity to certain populations, or differences in the allele frequency compared to other analyzed populations. For example, one insertion polymorphism at locus chr5-8948 would have been considered a fixed insertion if African-Americans were not included. Even though the insertion was common in African-Americans, two individuals also showed a heterozygous genotype. In other cases, the allele frequency varied notably between one population, most commonly African-Americans, and all other populations. Both types of distributions are displayed in Fig. 4, with the polymorphism alternately more or less common compared to the other three populations (*e.g.* 15% allele frequency in African-Americans versus more than 50% in the other populations).

Among the 32 L1 polymorphisms, eight insertions were absent in at least one of the four tested populations. For instance, the L1 insertion polymorphism of locus chr1-8645 was relatively common in Caucasians and South Americans (allele frequencies of 35% and 30%, respectively), but absent in Asians and African-Americans. Another insertion polymorphism (locus chr15-5300) was found only in the African-American population. Overall, a higher genetic diversity of the African-American population compared to the other three populations was evident. Polymorphisms were more often absent or very uncommon in the Asian population (34.4%) than in any other population (9.4% to 15.6%).

3.5. Sequence analysis of the polymorphic LINE-1 insertions

We identified five L1 insertions (15.6%) with 5-inversions. Five other insertions lacked a TSD; a TSD was considered present if 4 bp or more adjacent to the insertion site showed 100% identity. Three of these loci were identified in the CHG. Four out of five insertions without a TSD shared a small deletion of 1 to 4 bp at the place of the insertion. While most polymorphic L1 insertions were truncated (24 out of 32 loci), eight (25%) L1 insertions were full length (6 kb or more). One of these full-length polymorphic L1 elements was identified in the CHG; all others were retrieved from the PHG. The majority of full-length insertions were present in all four populations, and members of the intermediate or high allele

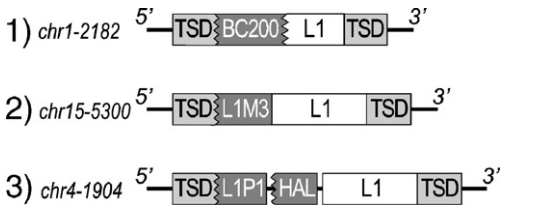

Fig. 5. Schematic structure of chimeric retroelements. All three chimeric L1 insertion polymorphisms are surrounded by TSDs between 15 and 18 bp. The chimeric retroelements are shown schematically; the original size of the insertion is not proportional to the size in this figure. A straight vertical line indicates that the element is not truncated; a sawtooth line, that the sequence is truncated. 1) Locus chr1-2182 with a BC200 sequence (truncated on both sides) directly adjacent to a truncated L1 insertion. 2) Locus chr15-5300 shows a full-length L1 element with another truncated L1 element. 3) Locus chr4-1904 with a full-length L1 element and two truncated elements (one L1, another L1-like) separated by short stretches of non-repetitive sequence.

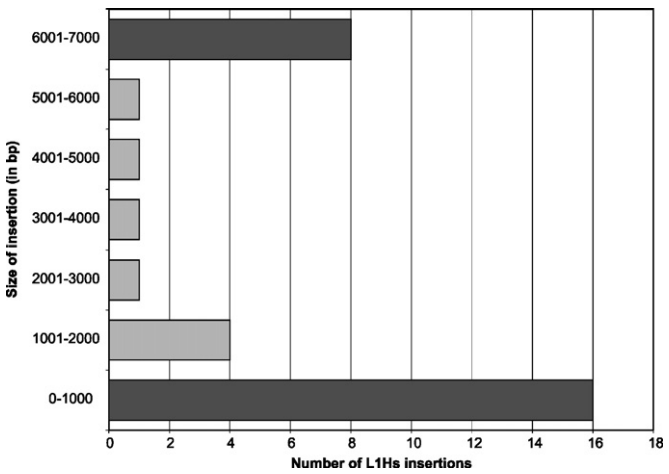

Fig. 6. Size distribution of polymorphic L1Hs insertions. L1Hs element size classes (in bp), showing the size distribution of insertions. Polymorphic L1 insertions showed a bias toward either full-length or short truncated elements (two bars in darker shade). 50% of the polymorphic insertions were 1000 bp or shorter, of which 11 insertions (34.4%) were less than 500 bp in length.

frequency classes of insertion polymorphisms. Locus chr5-8948 was identified from the CHG and had the highest allele frequency of all identified polymorphic L1 elements (see Fig. 4 and Table 2). Conversely, three full-length insertions were less common, and classified as insertions with low allele frequencies. Two of these three were not present in all populations. Locus chr4-1398 was completely absent from Caucasians and present in South Americans with a very low allele frequency. Locus chr15-5300 was present only in African-Americans, with an allele frequency of 27.5%, and absent in all other populations.

Two out of eight full-length L1 insertions had additional sequences adjacent to the L1Hs insertion at the 5-end. Altogether, three different loci exhibited extra sequences at the 5-end. By contrast, insertions with extra sequences adjacent to the 3-end were not found in this analysis. The 5-sequences showed considerable differences; *e.g.*, a fraction of L1P1 combined with a truncated HAL element, a BC200, and a truncated L1M3 (see Fig. 5). HAL is a LINE-like retroelement (Smit, 1999), while BC200 is a small non-messenger RNA mainly expressed in the nervous system (Tiedge et al., 1993). The additional sequences ranged in length from 141 to 226 bp. All three insertions were surrounded by TSDs with a size of 15 to 18 bp. However, each of these polymorphic insertions showed different characteristics. The insert at locus chr15-5300 also existed on chr1 in the PHG with a high level of sequence identity, and shared the same start of the truncated L1M3 element. The poly-A tail of the full-length L1Hs-Ta insert was much shorter compared to the poly-A tail of the L1 identified in our study. Also, the sequence on chr1 was not flanked by a TSD; and an *AluSx* element was immediately adjacent to the L1M3 element.

By comparison, the sequence at locus chr4-1904, with an L1P1, a HAL, and a full-length L1-Ta element in this chimeric structure, could not be identified a second time in the PHG. The repetitive sequences were separated by short stretches of non-repetitive sequences. The short sequence between L1P1 and HAL matched 100% on chr3, but no repetitive sequences were identified in the adjacent genomic region. The third insert at locus chr1-2182 differed in two ways from the two previously described loci; the L1-Ta insertion was truncated, and a 177 bp long BC-200 sequence was adjacent to the insertion.

The average size of all confirmed polymorphic L1 insertions was, at 2404 bp (ranging from 88 to 6065 bp), larger than the average of 1070 bp for all L1Hs insertions quoted in the literature (Lander et al., 2001). We did not define the average for Ta and pre-Ta elements, as there were too few elements present in our study to perform a meaningful analysis. In particular, the analysis would be biased towards larger sizes, as several truncated insertions were too small to be classified as Ta or pre-Ta elements. However, most of the pre-Ta insertions were relatively small (245–782 bp) with one exception: a full-length L1Hs-pre-Ta insertion identified in the PHG genome. In general, as Fig. 6 shows, the analyzed L1 insertions are biased toward very small (<1000 bp) or full-length L1 insertions (>6 kb). Altogether, these two groups account for ~75% of all confirmed polymorphic insertions.

4. Discussion

4.1. Computational identification of novel LINE-1 insertions

In this study, we have presented a comparison of two different assemblies of the human genome, the PHG and the CHG, toward identifying and characterizing novel L1 insertion polymorphisms. The underlying computational approach represents an efficient and accurate method to identify new polymorphic L1 insertions in the human genome. More than 94% of the selected loci were confirmed to be polymorphic. The high success rate and the absence of only one insertion from a survey of 80 diverse humans (160 chromosomes) indicate that the majority of the identified insertions are relatively common in the human population and represent intermediate frequency insertion polymorphisms. The low frequency insertion polymorphism may in fact be a private insertion, and in the context of population genetic studies will be of limited utility.

In the past, mobile element insertion polymorphisms have been identified via either computational or PCR-display based approaches. With these two methods, along with studies identifying the causes of certain diseases, approximately 342 published non-redundant polymorphic L1 insertions have previously been identified (Wang et al., 2006b). Both computational and PCR-display approaches have benefits and weaknesses. Using a computational data-mining analysis of the sequence database, insertions are identified with variable insertion frequencies, with a trend towards insertions with higher allele frequencies. A strength of this method is the relatively easy identification and determination of whether a given insertion has been previously identified. However, extracting L1 insertions from a single genome does not demonstrate if the insertion is polymorphic. Therefore, a PCR-based analysis is essential to identify the level of associated insertion polymorphism.

It is also important to note that insertions can only be identified if they are present in the genome database. Moreover, this approach is very much dependent on the accuracy of the reconstruction of the genome. By contrast, polymorphic insertions can be identified directly with a PCR-based display method. But this approach is biased toward low frequency, population specific or even private insertions, as insertions with higher allele frequencies would be present in multiple genomes and thus not be identified as polymorphic. One disadvantage of this method is the relatively high cost and amount of work for characterization of each element. In contrast, a comparison of two different draft genomes constitutes a relatively quick and less labor intensive method for identifying new polymorphic insertions. Not only are more insertions identified through the comparison of two different genomes compared to the use of one genome; additionally, there is usually a clear indication that a given locus is polymorphic.

It is commonly thought that older mobile elements are less likely to be polymorphic. This is illustrated with studies of L1Hs-Ta and pre-Ta elements, which were 45% and 14% polymorphic, respectively (Myers et al., 2002; Salem et al., 2003a). Thus, as the age of insertions increases, a computational approach based on a single genome becomes incrementally less

successful. By comparison, we identified five L1-pre-Ta insertions, all of which were polymorphic. In addition, 25 out of 26 identified L1Hs-Ta insertions, and all (six) L1Hs loci were confirmed as polymorphic (see Table 1). The L1PA2 element identified using our computational approach was not an authentic insertion polymorphism.

As disadvantages of computational comparisons of different genomes, success is dependent on the accuracy of both genome assemblies, and influenced by the genetic background and diversity of the genomic libraries that were sequenced to produce the genome assemblies. The assembly problem is a particular concern, as the assembly of genomic sequences is particularly difficult and error prone in areas of repetitive sequences. Another potential pitfall is the risk for incorrect alignments of paralogous repetitive sequences during the *in silico* analysis. This risk can be significantly reduced through careful inspection of insertions and flanking sequences before running a PCR-based analysis.

The computational comparison of genomic sequences from different individuals has proven successful for the identification of polymorphic mobile elements (Bennett et al., 2004; Wang et al., 2006a). Our study represents the first extraction of all novel polymorphic L1 insertion polymorphisms through the comparison of two different draft genome sequence assemblies, followed by a PCR-based population genetic diversity analysis of all potentially polymorphic loci. With the second part of our analysis, we not only confirm that the candidate loci are polymorphic, but also show the potential use of these insertions for future population genetic studies.

4.2. Population diversity

In 1973, Reed stated the ideal locus for distinguishing between two populations is one where an allele is fixed in one group but absent in the second (Reed, 1973). This criterion appears impossible to fulfill for human populations, as too little time has passed since the radiation of human population ~80,000 years ago. Additionally, several populations have experienced significant admixture at different times in history. Consequently, for population diversity studies, a set of many different markers like single nucleotide polymorphisms (SNPs) or mobile element insertion polymorphisms is useful (Bamshad et al., 2003; Ray et al., 2005b; Watkins et al., 2003). In comparison with other markers, mobile elements have two advantages: identity by descent with a very low likelihood of homoplasy and known ancestral state (Batzer and Deininger, 2002; Batzer et al., 1994; Ho et al., 2005). In contrast to *Alu* elements, absolutely no insertion homoplasy has previously been reported for LINE insertions. This is most likely based on the more complex nature of L1 insertions, such as variable L1 insertion length due to 5-truncations and somewhat lower retrotransposition rate (Ho et al., 2005; Salem et al., 2005).

With the identification and characterization of more polymorphic mobile elements including L1, the pool of population genetic markers will be increased. Thus, more specific markers better satisfying the requirements of population genetic studies can be identified. In this context, several novel L1 insertion polymorphisms have shown potentially useful diversity patterns for population genetic studies, such as varying allele frequencies between different populations and absence in some of the analyzed populations. The L1 polymorphism restricted to African-Americans represents one of few insertions that is not a very rare or even "private" insertion, but seems specific to a population or a subpopulation. Detailed characterization of the genetic diversity associated with this locus must be conducted to determine the specificity of this insertion. The elevated diversity over all polymorphic loci and higher combined allele frequency of African-Americans compared to the other three populations are consistent with previous findings that Africans have a greater genetic diversity compared to other populations (Armour et al., 1996; Bowcock et al., 1994; Cann et al., 1987; Hammer, 1995; Stoneking et al., 1997; Tishkoff et al., 1996; Vigilant et al., 1991). This can be attributed to a larger population size, a higher gene flow rate in the African-American population, and the origination of the human population in Africa (Stoneking et al., 1997).

Some polymorphism patterns of the Asian population are less easily explained, including the overall lower allele frequency and the absence of several insertions. We believe this might be caused by the genetic background of the two genomes being analyzed (*e.g.* Asians might be underrepresented in both genomes). However, the precise mechanism that generated this pattern of genetic diversity is unknown, as the exact genetic make up of the two genome assemblies is not defined.

4.3. Differences between L1 insertions in the CHG versus PHG

We believe the higher prevalence of occupied L1 insertion sites in the PHG (24 vs. 8 for novel insertions; 88 vs. 19 for the entire list of preliminary candidates) likely resulted from differences in the methods for creation and/or the genetic background of the two genomes. The CHG is more prone to gaps in areas of repetitive elements than the PHG due to its method of creation (shotgun vs. Bacterial Artificial Chromosome clones). Also, it appears likely that the CHG was derived from fewer diverse individuals than the PHG. The CHG was constructed from five individuals across four different geographic regions (Venter et al., 2001). For the PHG we know only that individuals from diverse populations close to the research facilities were randomly selected (Lander et al., 2001). It is also unclear how many diverse population groups were used at the time of creation for each genomic library. These factors, whether individually or in combination, may explain fewer occupied polymorphic L1 sites in the CHG than in the PHG.

4.4. Structural characteristics

The absence of TSDs in five cases, of which four insertion sites also showed a small deletion compared to the empty site, is likely caused either by genomic rearrangements or endonuclease independent insertion (Gilbert et al., 2002; Morrish et al., 2002; Myers et al., 2002). The frequency of inversed L1 insertions (15.6%) was lower or comparable to previous studies of L1Hs elements (23%, Ostertag and Kazazian, 2001). This difference is probably due to our study's small sample size. On the other hand, the frequency of young/new L1 insertions short

truncated or full-length insertions is in agreement with previous studies (Chen et al., 2005; Myers et al., 2002).

Polymorphic full-length L1 elements are of particular interest, as they have been inserted into the human genome recently. Particularly insertions with lower allele frequencies or presence in limited populations (*e.g.* locus chr15-5300) are potentially active driver genes. After a hot L1 insertion (from which most new insertions are derived) reaches an intermediate gene frequency in a population, it has usually accumulated a significant proportion of cool alleles throughout the population as a consequence of point mutations (Brouha et al., 2003; Lutz et al., 2003; Seleme et al., 2006). Thus, the retrotransposition activity of L1 insertions within a population can vary widely. Over time, some individuals/populations may still retain a highly active "hot" L1 insertion. Others would only have cool alleles of insertions; thus apart from the L1 insertion itself, they would have no further impact on the human genome.

4.5. Chimeric retroelements

For the three polymorphic L1 insertions with additional immediately adjacent 5-sequences, different models include 5-transduction, *in vivo* RNA–RNA hybridization, and recombination events (reviewed in Kazazian, 2004). While 3-L1-transduction occurs more frequently, 5-transduction plays a much smaller role (Goodier et al., 2000; Moran et al., 1999; Pickeral et al., 2000). In our computational approach, 3-transductions would not be identified, as flanking sequences matching both genomes and an adjacent poly-A tail of the L1 insert are required for identifying L1 polymorphisms. *In vivo* RNA–RNA hybridization can create chimeric retrotranscripts which are composed of full-sized copies of small nuclear RNA, fused at their 3-termini with 5-truncated, 3-poly-A tailed L1s (Buzdin et al., 2003). In both scenarios, the insertions are usually flanked by TSDs. In the third model, recombination events between L1 elements involving the TSDs are less likely to be identified.

Apart from the presence of TSDs around the three insertions, there is some evidence that different molecular mechanisms may have generated similar results. For the full-length L1 with a truncated L1M3 insert (locus chr15-5300), we found some indication that the underlying process could be 5-transduction. The insert is also identified on chr1 with a very high level of sequence identity, but with a shorter poly-A tail, no TSDs, and unique flanking sequences. In particular, the shorter poly-A tail may indicate that the insert on chr1 is older than the polymorphic L1 (Chen et al., 2005; Ovchinnikov et al., 2001; Roy-Engel et al., 2002b), suggesting that the sequence on chr1 was the originator of the newly identified polymorphic insertion. We have no clear explanation for the complicated structure of the insertion at locus chr4-1904. It could be that this insertion occurred through rearrangements including non-homologous recombination. Further, more detailed studies are necessary for the characterization of the underlying process.

The third locus (chr1-2182) differed from the other two in several respects, including the type of additional sequence (BC200) and a truncated L1Hs-Ta element. This complex element may be the result of RNA mediated sequence hybridization. Other known mechanisms for generating chimeric retroelements appear less likely for generating this hybrid structure.

5. Conclusions

Here, we have shown that the comparison of different human genome assemblies is a valuable, accurate, efficient method for identifying novel polymorphic L1 insertions (94% of selected polymorphic loci were confirmed polymorphic). This indicates that the analysis of human diversity generated by L1 insertions and other repetitive elements is far from complete. With the emergence of more diverse human genome sequences, many yet-unidentified L1 polymorphisms will be discovered. In particular, genomes covering a diverse population spectrum will give rise to additional mobile element insertion polymorphisms, many of which are likely to be population specific.

Acknowledgments

This research was supported by National Science Foundation BCS-0218338 and EPS-0346411 (MAB), Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05, (2000-05)-01 and (2001-06)-02 (MAB), National Institutes of Health RO1 GM59290 (MAB), R03 CA101515 (PL), P30 CA16056 (Roswell Park Cancer Institute), and the State of Louisiana Board of Regents Support Fund (MAB). We thank the whole Batzer lab, in particular Jianchuan Xing, Richard Cordaux, Jerilyn Walker, Hui Wang, and Scott Herke, as well as Lei Song in the Liang lab for their technical support, discussions, and comments on the manuscript. Special thanks also to Brygg Ullmer.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Armour, J.A.L., et al., 1996. Minisatellite diversity supports a recent African origin for modern humans. Nat. Genet. 13, 154–160.

Badge, R.M., Alisch, R.S., Moran, J.V., 2003. ATLAS: a system to selectively identify human-specific L1 insertions. Am. J. Hum. Genet. 72, 823–838.

Bamshad, M.J., Wooding, S., Watkins, W.S., Ostler, C.T., Batzer, M.A., Jordem, L.B., 2003. Human population genetic structure and inference of group membership. Am. J. Hum. Genet. 72, 578–589.

Batzer, M.A., Deininger, P.L., 1991. A human-specific subfamily of Alu sequences. Genomics 9, 481–487.

Batzer, M.A., Deininger, P.L., 2002. Alu repeats and human genomic diversity. Nat. Rev., Genet. 3, 370–379.

Batzer, M.A., Gudi, V.A., Mena, J.C., Foltz, D.W., Herrera, R.J., Deininger, P.L., 1991. Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res. 19, 3619–3623.

Batzer, M.A., et al., 1994. African origin of human-specific polymorphic Alu insertions. Proc. Natl. Acad. Sci. U. S. A. 91, 12288–12292.

Batzer, M.A., et al., 1996. Genetic variation of recent Alu insertions in human populations. J. Mol. Evol. 42, 22–29.

Bennett, E.A., Coleman, L.E., Tsui, C., Pittard, W.S., Devine, S.E., 2004. Natural genetic variation caused by transposable elements in humans. Genetics 168, 933–951.

Boissinot, S., Chevret, P., Furano, A.V., 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol. Biol. Evol. 17, 915–928.

Boissinot, S., Entezam, A., Young, L., Munson, P.J., Furano, A.V., 2004. The insertional history of an active family of L1 retrotransposons in humans. Genome Res. 14, 1221–1231.

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., Cavalli-Sforza, L.L., 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368, 455–457.

Brouha, B., et al., 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U. S. A. 100, 5280–5285.

Buzdin, A., et al., 2003. The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Res. 31, 4385–4390.

Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. Nature 325, 31–36.

Carroll, M.L., et al., 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J. Mol. Biol. 311, 17–40.

Chen, J.M., Chuzhanova, N., Stenson, P.D., Ferec, C., Cooper, D.N., 2005. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Human Mutat. 25, 207–221.

Deininger, P.L., Batzer, M.A., 1993. Evolution of retroposonsin. In: Hecht, M.K., et al. (Eds.), Evolutionary Biology. Plenum Press, New York, pp. 157–196.

Deininger, P.L., Batzer, M.A., 1999. Alu repeats and human disease. Mol. Genet. Metab. 67, 183–193.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., Kazazian Jr., H.H., 1991. Isolation of an active human transposable element. Science 254, 1805–1808.

Fanning, T., Singer, M., 1987. The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic Acids Res. 15, 2251–2260.

Feng, Q., Moran, J.V., Kazazian Jr., H.H., Boeke, J.D., 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87, 905–916.

Gilbert, N., Lutz-Prigge, S., Moran, J.V., 2002. Genomic deletions created upon LINE-1 retrotransposition. Cell 110, 315–325.

Goodier, J.L., Ostertag, E.M., Kazazian Jr., H.H., 2000. Transduction of 3′-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. 9, 653–657.

Hammer, M.F., 1995. A recent common ancestry for human Y chromosomes. Nature 378, 376–378.

Hillis, D.M., 1999. SINEs of the perfect character. Proc. Natl. Acad. Sci. U. S. A. 96, 9979–9981.

Ho, H.J., Ray, D.A., Salem, A.H., Myers, J.S., Batzer, M.A., 2005. Straightening out the LINEs: LINE-1 orthologous loci. Genomics 85, 201–207.

Istrail, S., et al., 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. PNAS 101, 1916–1921.

Jurka, J., 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. U. S. A. 94, 1872–1877.

Kazazian Jr., H.H., 1998. Mobile elements and disease. Curr. Opin. Genet. Dev. 8, 343–350.

Kazazian Jr., H.H., 2004. Mobile elements: drivers of genome evolution. Science 303, 1626–1632.

Kazazian Jr., H.H., Moran, J.V., 1998. The impact of L1 retrotransposons on the human genome. Nat. Genet. 19, 19–24.

Kazazian Jr., H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, S.E., 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332, 164–166.

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome. Res. 12, 656–664.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Lee, J., Cordaux, R. Han, K. Hedges, D. Liang, P., Batzer, M., in press. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. Gene. doi:10.1016/j.gene.2006.08.029.

Lutz, S.M., Vincent, B.J., Kazazian Jr., H.H., Batzer, M.A., Moran, J.V., 2003. Allelic heterogeneity in LINE-1 retrotransposition activity. Am. J. Hum. Genet. 73, 1431–1437.

Mathias, S.L., Scott, A.F., Kazazian Jr., H.H., Boeke, J.D., Gabriel, A., 1991. Reverse transcriptase encoded by a human transposable element. Science 254, 1808–1810.

Moran, J.V., DeBerardinis, R.J., Kazazian Jr., H.H., 1999. Exon shuffling by L1 retrotransposition. Science 283, 1530–1534.

Morrish, T.A., et al., 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat. Genet. 31, 159–165.

Myers, J.S., et al., 2002. A comprehensive analysis of recently integrated human Ta L1 elements. Am. J. Hum. Genet. 71, 312–326.

Nikaido, M., Rooney, A.P., Okada, N., 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales. Proc. Natl. Acad. Sci. U. S. A. 96, 10261–10266.

Ostertag, E.M., Kazazian Jr., H.H., 2001. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. 11, 2059–2065.

Ovchinnikov, I., Troxel, A.B., Swergold, G.D., 2001. Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res. 11, 2050–2058.

Perna, N.T., Batzer, M.A., Deininger, P.L., Stoneking, M., 1992. Alu insertion polymorphism: a new type of marker for human population studies. Hum. Biol. 64, 641–648.

Pickeral, O.K., Makalowski, W., Boguski, M.S., Boeke, J.D., 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. 10, 411–415.

Ray, D.A., et al., 2005a. Inference of human geographic origins using Alu insertion polymorphisms. Forensic Sci. Int. 153, 117–124.

Ray, D.A., et al., 2005b. Alu insertion loci and platyrrhine primate phylogeny. Mol. Phylogenet. Evol. 35, 117–126.

Ray, D.A., Xing, J. Salem, A.E. and Batzer. M.A., in press. SINEs of a nearly perfect character. Syst Biol.

Reed, T., 1973. Number of gene loci required for accurate estimation of ancestral population proportions in individual human hybrids. Nature 244, 575–576.

Roy-Engel, A.M., et al., 2001. Alu insertion polymorphisms for the study of human genomic diversity. Genetics 159, 279–290.

Roy-Engel, A.M., et al., 2002a. Non-traditional Alu evolution and primate genomic diversity. J. Mol. Biol. 316, 1033–1040.

Roy-Engel, A.M., et al., 2002b. Active Alu element "A-tails": size does matter. Genome Res. 12, 1333–1344.

Rozen, S., Skaletsky, H.J., 1998. Primer3Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Salem, A.H., Myers, J.S., Otieno, A.C., Watkins, W.S., Jorde, L.B., Batzer, M.A., 2003a. LINE-1 preTa elements in the human genome. J. Mol. Biol. 326, 1127–1146.

Salem, A.H., et al., 2003b. Alu elements and hominid phylogenetics. Proc. Natl. Acad. Sci. U. S. A. 100, 12787–12791.

Salem, A.H., Ray, D.A., Batzer, M.A., 2005. Identity by descent and DNA sequence variation of human SINE and LINE elements. Cytogenet. Genome Res. 108, 63–72.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467.

Sassaman, D.M., et al., 1997. Many human L1 elements are capable of retrotransposition. Nat. Genet. 16, 37–43.

Seleme, M.d.C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., Kazazian Jr., H.H., 2006. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. PNAS 103, 6611–6616.

Sheen, F.M., et al., 2000. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res. 10, 1496–1508.

Skowronski, J., Fanning, T.G., Singer, M.F., 1988. Unit-length line-1 transcripts in human teratocarcinoma cells. Mol. Cell. Biol. 8, 1385–1397.

Smit, A.F., 1996. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 6, 743–748.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Smit, A.F., Toth, G., Riggs, A.D., Jurka, J., 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J. Mol. Biol. 246, 401–417.

Smit, A., Hubley, R., Green, P., 1996. RepeatMasker Open-3.0.

Stoneking, M., et al., 1997. Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res. 7, 1061–1071.

Tiedge, H., Chen, W., Brosius, J., 1993. Primary structure, neural-specific expression, and dendritic location of human BC200 RNA. J. Neurosci. 13, 2382–2390.

Tishkoff, S.A., et al., 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271, 1380–1387.

Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., Wilson, A.C., 1991. African populations and the evolution of human mitochondrial DNA. Science 253, 1503–1507.

Wang, J., et al., 2006a. Whole genome computational comparative genomics: a fruitful approach for ascertaining Alu insertion polymorphisms. Gene 365, 11–20.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M., Liang, P., 2006b. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Human Mutat. 27, 323–329.

Watkins, W.S., et al., 2003. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Res. 13, 1607–1618.